
Identifying User Survival Types via Clustering of Censored Social Network Data

S Chandra Mouli¹ Abhishek Naik² Bruno Ribeiro¹ Jennifer Neville¹

¹*Purdue University*

²*Indian Institute of Technology Madras*

Abstract

The goal of cluster analysis in survival data is to identify clusters that are decidedly associated with the survival outcome. Previous research has explored this problem primarily in the medical domain with relatively small datasets, but the need for such a clustering methodology could arise in other domains with large datasets, such as social networks. Concretely, we wish to identify different survival classes in a social network by clustering the users based on their lifespan in the network. In this paper, we propose a decision tree based algorithm that uses a global normalization of p -values to identify clusters with significantly different survival distributions. We evaluate the clusters from our model with the help of a simple survival prediction task and show that our model outperforms other competing methods.

1 Introduction

Survival analysis is used to model the length of time until a particular event occurs, for example, the time until death of a medical patient or failure of an equipment [19]. One of the tasks in survival analysis is to cluster the individuals in a semi-supervised fashion by using not only their attributes but also their survival times. In other words, the goal of this task is to group the individuals with similar survival times into a single cluster. For example, the individuals could be divided into ‘high-risk’ and ‘low-risk’ groups, assuming there are only two clusters.

There are numerous applications for such a clustering procedure, one of which is the identification of cancer subtypes from gene expression data and is the most studied in the literature. Another example is that of grouping users based on their survival in a social network (i.e., the time until they leave the system permanently). Such an analysis can be extremely valuable as it can help in categorizing new users into groups that provide key information about their survival times.

There have been previous attempts at an unsupervised approach to clustering where the clusters were identified by considering only the attributes and not the survival outcome [10, 2, 6]. This approach has a clear drawback that the clusters obtained may be completely unrelated to the survival of these grouped individuals. A second approach to clustering is to divide the individuals purely based on their survival times [26, 31]. But these approaches do not provide us with any meaningful information about the connection between the features and the survival outcome. There also have been several semi-supervised clustering approaches [1, 4, 5, 20] proposed that use some

form of label information to obtain the clusters, but most of these methods do not work well when the data is censored, which is a common characteristic of survival data.

Only a few methods have been proposed that perform supervised clustering on censored data. Recently, Gaynor and Bair [11] proposed a supervised version of the sparse clustering algorithm [33]. Sparse clustering provides a technique for feature selection in clustering by assigning weights to each feature. Supervised sparse clustering simply alters the initial weights of the features to reflect the features' relative importance in predicting survival.

We note that the major expectation from the resultant clusters is that they have considerably different survival distributions (Section 2.1). In this paper, we utilize this fact and propose a novel partially supervised clustering approach for survival data. We predominantly work with social network data and obtain clusters of users based on their survival in the system.

2 Preliminaries

In what follows, we describe the important concepts relating to survival analysis that are used in this paper.

2.1 Survival Distribution & Hazard Function

Survival distribution is defined as the probability that an individual survives atleast until time t , and is given by

$$S(t) = P(T > t) = 1 - F(t), \quad 0 < t < \infty, \quad (2.1)$$

where T is a nonnegative random variable representing the time of death of an individual, and $F(t)$ is the cumulative distribution function. In survival applications, it is typically convenient to define the hazard function, that represents the instantaneous rate of death of an individual given that she has survived till time t . The hazard function $\lambda(t)$, is given by

$$\lambda(t) = \frac{f(t)}{S(t)}, \quad (2.2)$$

where $f(t)$ is the probability density function.

2.2 Right Censoring

When working with survival times of individuals, it is common to have censored observations. This happens when the event in consideration does not occur until end of the study. Consider, for example, the time until a user in a social network stops being active. In this scenario, we say the observation of a user is (right) censored if, at the time of data collection, the user is still active, i.e., the death event has not yet occurred. It is clear that ignoring the effect of censoring can lead to skewed estimates of survival probabilities. Right censoring can be classified into three types, namely, Type-I, Type-II and Random censoring [19]. Random censoring is a common feature when the individuals enter the study at different times, which is notably the case in the social network scenario, where the users join the system at different times (Figure 5.1). Here, we assume that the censoring times are independent of the death times, which is justified when the joining times are random [19]. In the following subsection, we define Kaplan-Meier estimator (or product limit estimator) that provides a method for incorporating the censoring effect while obtaining the survival probabilities.

2.3 Kaplan-Meier Estimator

Kaplan-Meier estimator [14] has been widely used in a variety of survival analysis tasks since its introduction. It provides a non-parametric maximum likelihood estimate of the empirical survival distribution, given by,

$$\hat{F}(t) = \prod_{j|t_j \leq t} \frac{n_j - d_j}{n_j}, \quad (2.3)$$

where d_j is the number of individuals who ‘die’ at time t_j and n_j is the number of individuals at risk of ‘death’ at time just prior to t_j , i.e., the individuals that are not ‘dead’ and not yet been censored.

3 Related Work

Majority of work in survival analysis has dealt with the task of predicting the survival outcome especially when the number of features is much higher than the number of subjects [34, 8, 12, 27]. A number of approaches have also been proposed to perform feature selection in survival data [13, 16]. In the social network scenario, Sun et al. [30] tried to predict the relationship building time, that is, the time until a particular link is formed in the network. They use generalized linear models [18] with a modified likelihood function that incorporates censoring.

There have been relatively fewer works that perform clustering on survival data. Many unsupervised approaches have been proposed to identify cancer subtypes in gene expression data [10, 2, 6]. However, we are interested in the task of supervised clustering for survival data. Traditional semi-supervised clustering methods [1, 4, 5, 20] do not perform well in this scenario since they do not provide a way to handle the issues with right censoring. Bair and Tibshirani [3] proposed a semi-supervised method for clustering survival data in which they assign Cox scores [9] for each feature (or gene) in their dataset and considered only the features with scores above a predetermined threshold. Then, an unsupervised clustering algorithm, like k-means, is used to group the individuals using only the selected features. Such an approach can miss out on clusters when they are weakly associated with the survival outcome since such features are discarded immediately after the initial screening.

In order to overcome this issue, Gaynor and Bair [11] proposed supervised sparse clustering as a modification to the sparse clustering algorithm of Witten and Tibshirani [33]. The sparse clustering algorithm uses an objective function similar to k-means but with the modification that each feature has a weight associated to it. Supervised sparse clustering [11] initializes these feature weights depending on the feature’s relation with the survival outcome and optimizes the same objective function. Once again, they use Cox scores [9] to quantify the effect of each feature on the survival outcome. The authors show that this leads to a clustering that is relatively more linked to the survival outcome.

Both of these methods have been shown to perform well when the dataset size is small. Supervised sparse clustering in particular, is computationally expensive since in each iteration, it performs an unsupervised k-means clustering over the entire dataset. In this paper, we propose a decision tree based clustering algorithm that not only identifies better clusters than the existing methods but can also work efficiently with large amounts of data.

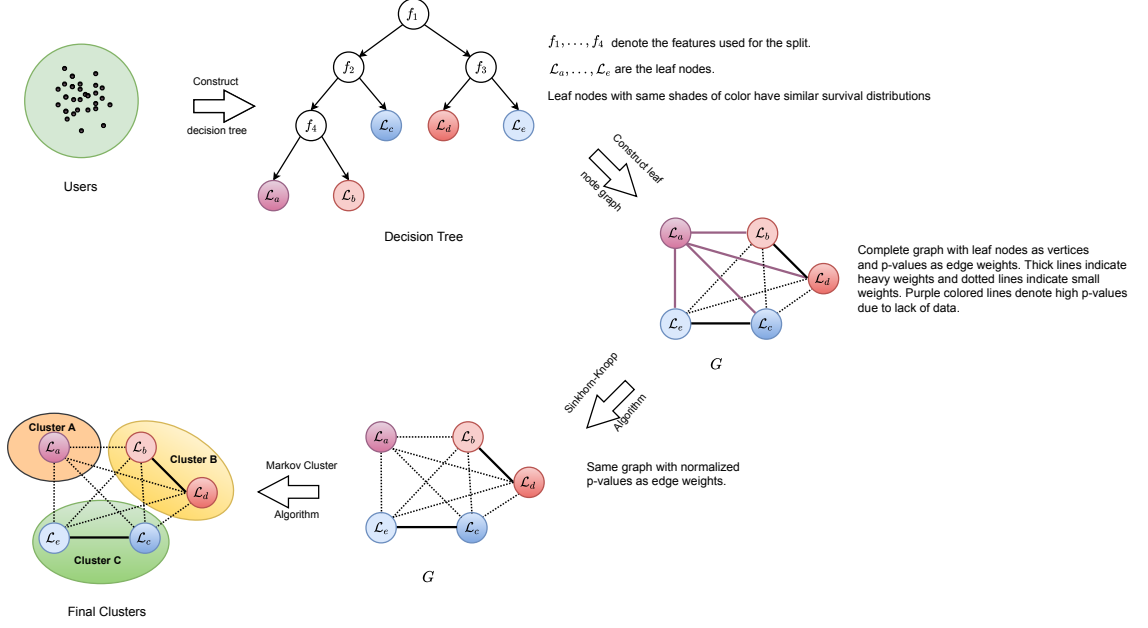


Figure 4.1: Image depicting the complete procedure to obtain the clusters.

4 Methodology

Our primary goal for clustering is that the survival distributions be different across clusters. In this section, we present a decision tree based approach that is built to optimize for this goal. The principal idea is to construct a decision tree such that the survival distributions of the two populations of users at each split differ significantly from each other. Concretely, we split the current set of users based on an attribute-value test and obtain the survival distributions of the two populations of users using Kaplan-Meier estimates (Section 2.3). We use Kuiper statistic [15] in order to quantify how significantly these survival distributions differ. This process is repeated for all attribute-value pairs and the one that results in the lowest p-value (denoting that the survival distributions after the split are most likely different from each other) is used as a node in the decision tree. The significance level, α , is a parameter to our algorithm (set at 0.05 in our experiments). It is important to note that we are performing many statistical tests at each node which leads to a multiple hypothesis test problem [24, 25]. We use the Bonferroni correction [24] to compensate for doing m statistical tests which reduces the significance level by a factor of m . Thus, a node is split only if the resultant p-value is below the corrected significance level α/m .

This procedure results in a tree where each leaf node has an associated population of users and thus, the leaf nodes themselves can be interpreted as clusters. But, the issue here is that the leaf nodes need not have significantly different distributions from each other. It is not hard to imagine that two leaf nodes descending from different parts of the tree may have very similar survival distributions. Hence, it is necessary to group these leaf nodes such that the ones with similar distributions are clustered together. Note that growing a tree deep and clustering the leaf nodes is different from growing a shallow tree and using the leaf nodes as clusters.

Let \mathcal{L} be the set of leaf nodes. In order to cluster these leaf nodes, we build a complete graph $G = (V, E)$, where $V = \mathcal{L}$ and $E = \{(i, j) : i, j \in \mathcal{L}\}$. Define a weighted adjacency matrix

$W \in \mathbb{R}_{\geq 0}^{|V| \times |V|}$ such that $W_{ij} = K_p(i, j)$ where $K_p : \mathcal{L} \times \mathcal{L} \rightarrow \mathbb{R}$ is a function that, given two leaf nodes, returns the p-value associated to the Kuiper’s test between the survival distributions of the two leaf nodes. In other words, the weights on the edge represent the degree of similarity between the survival distribution of the two vertices. Now we could perform a graph clustering procedure on G in order to cluster the leaf nodes. However, using p-values directly as edge weights is not a sound approach. This is because p-values can be high for two reasons – the distributions in question are very similar, or there is not enough data to significantly claim that the distributions differ from each other. Thus, a leaf node with very few associated users will form heavy edges with all the other leaf nodes, resulting in a clustering with just one group. We normalize the weight matrix W using Sinkhorn-Knopp algorithm [28] that converts it into a doubly stochastic matrix W_{sk} , thereby solving the aforementioned issue. We use Markov cluster algorithm [32] on the graph G and weight matrix W_{sk} in order to obtain a clustering of leaf nodes and consequently, a clustering on the entire set of users.

5 Dataset

In this paper, we analyze a large-scale social network dataset collected from Friendster. Friendster was founded in 2002 and was one of the earliest social networking websites, reaching 3 million users within the first few months [23]. The website allowed users to share messages, photos and videos with other members. Each user also had a profile page consisting of general information like name, gender, age, location and interests.

After processing 30TB of data, originally collected by the Internet Archive in June 2011, the resulting network has around 15 million users with 335 million friendship links. Each user has profile information such as age, gender, and marital status. Additionally, there are user comments on each other’s profile pages with timestamps that indicate activity in the site. See Table 6.1 for some additional statistics on the dataset.

Number of Users	15M
Number of friendship links	335M
Number of comments	75M
Number of users with atleast one comment	9.5M
Number of users with atleast ten comments	1.93M
Number of users with Age, Gender & Location specified	6.47M

Table 5.1: Statistics on the Friendster dataset

Since, we do not have users’ login information, we use the comments sent and received by the users as a proxy for activity. We choose ten months of inactivity to be the cut-off period after which the user will be assumed to have left the social network. The time from the user’s joining to her last comment will be considered as her lifetime in the system.

Ribeiro and Faloutsos [22] depicted the effect of the introduction of “new Facebook wall” in July 2008 to other competing social networking websites including Friendster. It is clear from their analysis that Friendster faced a continuous decline in the number of daily active users since then. Seeing that we wish to analyze the system on its own without any external influence, we only use the data upto March 2008 (six years from the introduction of Friendster) and disregard the rest. Figure 5.2 shows the estimated survival distributions for the entire data and the reduced data. Note the sudden drop in survival probabilities when using the complete data, which is missing when we use only the data prior to the introduction of “new Facebook wall”. In this work, we only consider

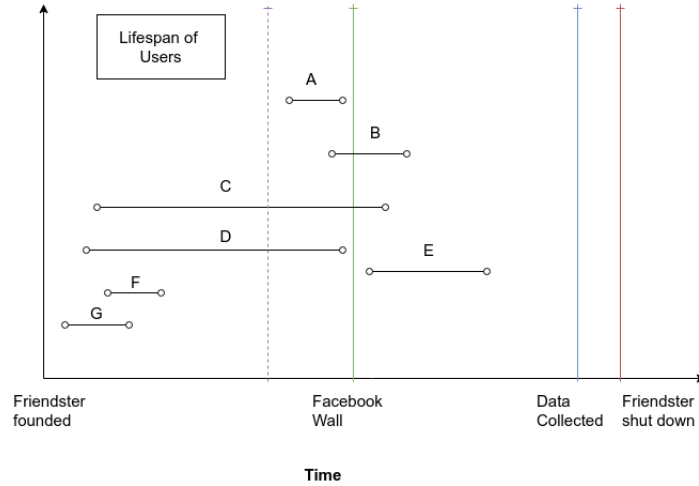


Figure 5.1: Image depicting the lifespan of users in Friendster when comments are used as a proxy for activity. The vertical dotted line indicates the cut-off period of ten months. Users with no activity in this period are considered ‘dead’. In this figure, users A, B, C & D have censored observations, user E is discarded and users F & G have known survival lifetime.

a subset of 1.2 million users who had participated in atleast one comment, had specified their age and gender, and had joined the social network before March 2008. Our processed data will be made available to the public once we get the RIB approval for its distribution.

6 Evaluation Metrics

In this section, we describe a survival prediction task that we designed to evaluate the quality of the clusters obtained from our procedure. We also describe briefly, two other standard evaluation techniques used in the literature, namely, hazard ratio and log rank test [29, 17].

Classification task

In order to validate our claim that the clusters obtained differentiate the users based on their survival outcome, we devise a classification task as follows - *given a user’s profile and activity information for the initial five months, predict whether she will stay in the system five months hence.*

We obtain the clusters from running different clustering procedures on the features generated from the initial five months’ data. We then use only these cluster labels as features in a logistic regression model [18] to predict whether the user will survive the next five months. A high prediction accuracy indicates that the clustering has extracted the information about the survival outcome from the entire set of features.

Hazard ratio

Hazard ratio is defined as the ratio of the hazard rates (Section 2.1) of two groups of entities. The Cox proportional hazards model [9] provides a method to estimate the hazard ratio given that the hazard ratio is constant over time. Spruance et al. [29] give a description of the interpretation and the correct usage of the hazard ratios.

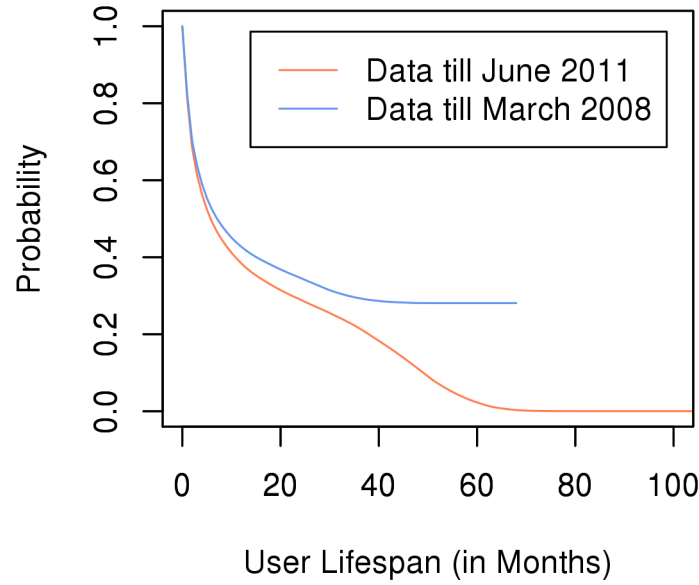


Figure 5.2: Survival distributions of the complete data and the reduced data.

Log-rank test

Log-rank test [17, 21] is a non-parametric hypothesis test that is widely used to compare two survival distributions. It tests the null hypothesis that the two (or more) groups in consideration have the same survival distributions. The predominant reason for the popularity of this test while comparing survival distributions is that it incorporates the effect of censoring the same way as the Kaplan-Meier estimates [7].

7 Results

We compare our model with two other clustering approaches – semi-supervised clustering of Bair and Tibshirani [3], and supervised sparse clustering by Gaynor and Bair [11]. We use user’s profile features (like age, gender, relationship status, occupation, location) as well as construct features based on the user’s initial five months’ activity (like number of comments sent and received, number of individuals interacted with, etc.). Since our model is based on decision trees, it can handle both categorical and numerical features with ease. In order for other clustering approaches to work effectively, we encode the categorical features like location using g binary variables, where g is the number of values the feature can take. Out of a total of 500 features, we choose 25 ($\approx \sqrt{500}$) top features found using Cox scores [9]. Semi-supervised clustering uses only these top features to find the clusters. Supervised sparse clustering assigns positive weights to these features and zero weights to the rest, and runs the standard sparse clustering algorithm with these initial weights.

Table 7.1 shows the values for the log-rank test and the hazard ratio for the clusters obtained from different clustering algorithms. The number of clusters were kept fixed at two in this exper-

Clustering Method	Log Rank Test (χ^2)	Hazard Ratio
Proposed Method	172557	3.242
Semi-Supervised Clustering [3]	141206	2.274
Supervised Sparse Clustering [11]	140660	3.331

Table 7.1: Log Rank test and Hazard ratio values for $k = 2$

	Precision	Recall	F-measure	Accuracy	FPR
Proposed Method ($k = 2$)	0.689	0.707	0.698	0.673	0.366
Proposed Method ($k = 3$)	0.711	0.612	0.658	0.659	0.286
Proposed Method ($k = 4$)	0.688	0.658	0.673	0.657	0.343
Proposed Method ($k = 5$)	0.689	0.665	0.677	0.660	0.345
Semi-Supervised Clustering ($k = 2$)	0.763	0.502	0.605	0.650	0.178
Semi-Supervised Clustering ($k = 3$)	0.730	0.584	0.649	0.662	0.249
Semi-Supervised Clustering ($k = 4$)	0.511	0.822	0.630	0.484	0.906
Semi-Supervised Clustering ($k = 5$)	0.727	0.591	0.652	0.662	0.255
Supervised Sparse Clustering ($k = 2$)	0.764	0.499	0.604	0.649	0.176
Supervised Sparse Clustering ($k = 3$)	0.732	0.579	0.647	0.661	0.243
Supervised Sparse Clustering ($k = 4$)	0.772	0.479	0.591	0.645	0.162
Supervised Sparse Clustering ($k = 5$)	0.725	0.509	0.598	0.633	0.222

Table 7.2: Classification results with features from various clustering algorithms for number of clusters, $k = 2, 3, 4, 5$. The clusters obtained from the proposed method achieve better accuracies and f-scores when $k = 2$ and 4 whereas the accuracies are comparable for $k = 3$. Highest accuracy across all algorithms is achieved by the proposed method when $k = 2$, that is, when there are only two classes of Friendster users: short-lived and long-lived.

iment. The χ^2 values shown in the table are huge, indicating that all three clustering algorithms return clusters that have significantly different distributions. The hazard ratios of clusters from our model and that from supervised sparse clustering are comparable.

The performance of logistic regression model using only the cluster labels as features is presented in Table 7.2 for the three clustering algorithms. We repeat the task for different values for k , the number of clusters. The clusters from our model have higher prediction accuracy than clusters from other models for $k = 2$ and 4 whereas the accuracy is comparable for $k = 3$ and 5. Our method also has higher f-measure scores compared to the competing models regardless of k .

8 Conclusion

In this paper, we proposed a partially supervised approach for clustering users based on their survival outcome. We used decision trees to divide the users such that the survival distributions of the subgroups are significantly different at each step. We then performed graph clustering over these subgroups in order to make sure that the subgroups with similar survival distributions are clustered together. Explicitly working with survival distributions effectively leads to a clustering that is highly associated with the survival outcome. We used our model in a social network dataset to identify groups of users with different survival types. We evaluated our model using two standard metrics, log-rank test and hazard ratio, and a classification task that we devised to measure the clusters' ability to predict survival. We also observed in our dataset that the classification accuracy

is highest when we use the proposed method to cluster the users into two groups – *short-lived* and *long-lived*.

References

- [1] Charu C Aggarwal, Stephen C Gates, and Philip S Yu. On using partial supervision for text categorization. *IEEE Transactions on Knowledge and data Engineering*, 16(2):245–255, 2004.
- [2] Ash A Alizadeh, Michael B Eisen, R Eric Davis, Chi Ma, Izidore S Lossos, Andreas Rosenwald, Jennifer C Boldrick, Hajeer Sabet, Truc Tran, Xin Yu, et al. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, 2000.
- [3] Eric Bair and Robert Tibshirani. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol*, 2(4):e108, 2004.
- [4] Sugato Basu, Arindam Banerjee, and Raymond Mooney. Semi-supervised clustering by seeding. In *In Proceedings of 19th International Conference on Machine Learning (ICML-2002)*. Citeseer, 2002.
- [5] Sugato Basu, Mikhail Bilenko, and Raymond J Mooney. A probabilistic framework for semi-supervised clustering. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 59–68. ACM, 2004.
- [6] Arindam Bhattacharjee, William G Richards, Jane Staunton, Cheng Li, Stefano Monti, Priya Vasa, Christine Ladd, Javad Beheshti, Raphael Bueno, Michael Gillette, et al. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences*, 98(24):13790–13795, 2001.
- [7] J Martin Bland and Douglas G Altman. The logrank test. *Bmj*, 328(7447):1073, 2004.
- [8] Hege M Bøvelstad, Ståle Nygård, Hege L Størvold, Magne Aldrin, Ørnulf Borgan, Arnoldo Frigessi, and Ole Christian Lingjærde. Predicting survival from microarray data: a comparative study. *Bioinformatics*, 23(16):2080–2087, 2007.
- [9] David R Cox. Regression models and life-tables. In *Breakthroughs in statistics*, pages 527–541. Springer, 1992.
- [10] Michael B Eisen, Paul T Spellman, Patrick O Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998.
- [11] Sheila Gaynor and Eric Bair. Identification of biologically relevant subtypes via preweighted sparse clustering. *Biostatistics*, pages 1–33, 2013.
- [12] Torsten Hothorn, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro, and Mark J Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2006.
- [13] Hemant Ishwaran, Udaya B Kogalur, Eiran Z Gorodeski, Andy J Minn, and Michael S Lauer. High-dimensional variable selection for survival data. *Journal of the American Statistical Association*, 105(489):205–217, 2010.

- [14] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.
- [15] Nicolaas H Kuiper. Tests concerning random points on a circle. In *Indagationes Mathematicae (Proceedings)*, volume 63, pages 38–47. Elsevier, 1960.
- [16] Vincenzo Lagani and Ioannis Tsamardinos. Structure-based variable selection for survival data. *Bioinformatics*, 26(15):1887–1894, 2010.
- [17] Nathan Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer chemotherapy reports. Part 1*, 50(3):163–170, 1966.
- [18] Peter McCullagh. Generalized linear models. *European Journal of Operational Research*, 16(3):285–292, 1984.
- [19] Rupert G Miller Jr. *Survival analysis*, volume 66. John Wiley & Sons, 2011.
- [20] Kamal Nigam, Andrew McCallum, Sebastian Thrun, Tom Mitchell, et al. Learning to classify text from labeled and unlabeled documents. *AAAI/IAAI*, 792, 1998.
- [21] Richard Peto and Julian Peto. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, pages 185–207, 1972.
- [22] Bruno Ribeiro and Christos Faloutsos. Modeling website popularity competition in the attention-activity marketplace. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 389–398. ACM, 2015.
- [23] Gary Rivlin. Wallflower at the web party. *New York Times*, 15(10), 2006.
- [24] G Rupert Jr et al. *Simultaneous statistical inference*. Springer Science & Business Media, 2012.
- [25] Juliet Popper Shaffer. Multiple hypothesis testing. *Annual review of psychology*, 46(1):561–584, 1995.
- [26] Margaret A Shipp, Ken N Ross, Pablo Tamayo, Andrew P Weng, Jeffery L Kutok, Ricardo CT Aguiar, Michelle Gaasenbeek, Michael Angelo, Michael Reich, Geraldine S Pinkus, et al. Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine*, 8(1):68–74, 2002.
- [27] Pannagadatta K Shivaswamy, Wei Chu, and Martin Jansche. A support vector approach to censored targets. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 655–660. IEEE, 2007.
- [28] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.
- [29] Spotswood L Spruance, Julia E Reid, Michael Grace, and Matthew Samore. Hazard ratio in clinical trials. *Antimicrobial agents and chemotherapy*, 48(8):2787–2792, 2004.
- [30] Yizhou Sun, Jiawei Han, Charu C Aggarwal, and Nitesh V Chawla. When will it happen?: relationship prediction in heterogeneous information networks. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 663–672. ACM, 2012.

- [31] Marc J Van De Vijver, Yudong D He, Laura J Van't Veer, Hongyue Dai, Augustinus AM Hart, Dorien W Voskuil, George J Schreiber, Johannes L Peterse, Chris Roberts, Matthew J Marton, et al. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, 2002.
- [32] Stijn Marinus Van Dongen. *Graph clustering by flow simulation*. PhD thesis, 2001.
- [33] Daniela M Witten and Robert Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010.
- [34] Daniela M Witten and Robert Tibshirani. Survival analysis with high-dimensional covariates. *Statistical methods in medical research*, 19(1):29–51, 2010.